Published in final edited form as:

Med Decis Making. 2011 ; 31(2): 270–280. doi:10.1177/0272989X10380924.

A Longitudinal Comparison of Five Preference-weighted Health State Classification Systems in Persons with Intervertebral Disc Herniation

Christine M. McDonough, PhD, PT^{1,6}, Tor D. Tosteson, ScD^{1,3}, Anna N.A. Tosteson, ScD^{1,2,3}, Alan M. Jette, PhD, MPH, PT⁶, Margaret R. Grove¹, and James N. Weinstein, DO, MS^{1,4,5}

¹The Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth College, Lebanon, NH

²Department of Medicine, Dartmouth-Hitchcock Medical Center, Lebanon NH

³Department of Community and Family Medicine, Dartmouth Medical School, Hanover, NH

⁴Department of Orthopaedics, Dartmouth Medical School, Hanover NH

⁵Dartmouth-Hitchcock Medical Center, Lebanon NH

⁶Health & Disability Research Institute, Boston University School of Public Health, Boston, MA

Keywords

Health State Preferences, Utilities, and Valuation; Health status indicators; Spine diseases; Quality of life; Economic evaluation; SPORT; Scale Validation; Cost-utility Analysis; Cost-effectiveness Analysis

Introduction

There is evidence of increasing use of cost-utility analysis to assess the relative value of alternative treatment interventions when resources are limited.[1, 2] To estimate Quality-Adjusted Life Years (QALYs) for the denominator of the incremental cost-effectiveness ratio (ICER), outcomes of treatment are measured using a single score, anchored at 0 for death and 1 for perfect health, and weighted for the relative desirability of the health state. Standards for economic evaluations recommend using societal values (utilities or preferences).[3] The two main approaches to obtaining “societal health state values” include: 1) direct measurement of values for health states of a representative sample of the population using methods such as standard gamble, time tradeoff, and visual analogue scale ratings, and 2) indirect measurement using preference-weighted health state classification systems such as the Quality of Well Being Scale,[4] the EuroQoL EQ-5D, [5, 6] the McMaster Health Utilities Index HUI,[7–9] or the SF-6D. [10, 11] In addition, methods have been developed to estimate health state values from existing HRQOL data, for example using regression models.

Preference-weighted health state classification systems are increasingly used in cost-utility analyses to estimate change in QALYs.[12] Furthermore, they are increasingly used as

Corresponding Author: Christine M. McDonough, the Dartmouth Institute for Health Policy and Clinical Practice, 565 Rubin Clinical Research Section, Dartmouth-Hitchcock Medical Center One Medical Center Drive, Lebanon, NH 03756. (603) 653-3565, christine.m.mcdonough@dartmouth.edu.

Commercial Support/Conflicts Statement: No conflicts to declare

measures of health outcome in clinical trials. Systems vary in their approaches to the design of each component: the descriptive system, preference measurement method, source of community preferences, and approaches to scoring. Questions remain about the comparability of systems in specific populations and about the extent to which differences in systems could impact the results of cost-utility analyses and therefore, policy decisions. In choosing among measurement systems researchers need to know the strengths and weaknesses of alternatives to optimize measurement performance for the particular problem under study, to interpret score changes or differences, and for study planning.

Evidence from cross-sectional comparisons indicates that significant variation exists in mean scores obtained from different systems. [13–19] However, when the research purpose is to measure change due to treatment, as in cost-utility analysis, longitudinal studies are necessary to evaluate system performance. Longitudinal head-to-head comparisons of preference-weighted systems indicate that change scores vary across systems used to estimate health state values (HSVs). [20–34] In particular, there is some evidence of difficulty measuring change at the worst levels of health (floor effects) for SF-derived systems,[15, 27, 35] and at the best levels (ceiling effects) for EQ-5D[14, 36] and in detecting changes in function unrelated to the extremities for HUI3.[8]

To our knowledge, there are no longitudinal comparisons of these systems in the population of patients with intervertebral disc herniation, a common and costly spine disorder. Therefore, we aimed to conduct a critical comparison of the measurement characteristics of the EQ-5D-UK, EQ-5D-US, HUI3, HUI2, SF-6D and an algorithm to estimate the QWB from SF-36 data (eQWB) among Spine Patient Outcomes Research Trial (SPORT) participants with intervertebral disc herniation (IDH).

Methods

Sample and selection

We used baseline and one year data from an ongoing prospective study of interventions for symptomatic lumbar spine disorders (SPORT). The design of this study has been previously reported in detail.[37] In brief, SPORT is a multi-center study including three randomized trials and three observational cohorts. To be eligible for SPORT, participants were 18 years or older and had a diagnosis of IDH, Spinal Stenosis (SpS), or Degenerative Spondylolisthesis (DS). Participants were excluded if there was evidence of non-surgical treatment for fewer than six weeks for IDH and twelve weeks for SpS and DS; cauda equina syndrome; contraindications to spine surgery; possible pregnancy; active malignancy; current fracture; infection; or prior lumbar spine surgery.[37]

Measures of Health State Value

The instruments used to characterize health state values are described below.

EQ-5D—The EuroQoL EQ-5D includes five attributes rated on three levels to define 245 health states (when “dead” and ‘unconscious’ are added). Using the same EQ-5D health state classification system with the reference time frame “today,” we applied EQ-5D-UK preference weights and EQ-5D-US weights. The UK (York) weights were measured using time-tradeoff values for a subset of health states from a sample of the UK population. [5, 6] The US weights were measured using time tradeoff in a representative sample of the US population.[38] Both systems use additive models of attribute independence with different adjustments for any health state at the worst possible level.

Health Utilities Index—The McMaster Health Utilities Index has been well described. [8, 9, 39–41] Using the same health state classification system, SPORT is licensed to apply the Mark 2 (HUI2) and the Mark 3 (HUI3) utility functions. HUI(2) represents seven attributes on four or five levels and defines 24,000 health states. HUI(3) has five or six levels for each of its eight attributes and encompasses 972,000 unique health states. The HUI(2) and HUI(3) use multiplicative multi-attribute utility functions based on visual analogue and standard gamble scores obtained from community samples in Canada. [9, 40, 41] The reference time frame for the questionnaire was “the past four weeks” and we did not include the fertility dimension in our survey.

SF-36-Derived Measures—The SF-6D, version 2, provides a method for deriving a preference score from the SF-36 instrument. [10, 11] It represents six attributes on up to six levels. An additive model was used and community weights were derived using standard gamble utilities from a UK population for a subset of health states.

Estimated Quality of Well-Being score (eQWB): The Quality of Well-Being scale (QWB) is a preference-based health measure that includes three additive functional dimensions and a symptom dimension.[4] Community preferences were measured using category rating for a representative sample of 866 adults in the San Diego area. Scores can range from 0.0 to 1.0, though the lowest score for a health state other than death is 0.32.[42] We previously estimated QWB scores using a regression model based on five subscales of the SF-36 reported by the Beaver Dam Health Outcomes Study. [42]

Criterion Measures

Criteria for comparison used for this study included a disease-specific measure, the Oswestry Disability Index (ODI) and patient ratings of satisfaction with symptoms (symptom satisfaction), self-perceived progress (progress rating), and self-perceived health (SPH).

Oswestry Disability Index(AAOS - modified version).[43–45]—The ODI includes nine items on six levels and yields an index score from least to most disability of 0 to 100. For consistency of interpretation, we subtracted ODI scores from 100 so that higher scores indicate better health.

Symptom Satisfaction—The participant is asked, "If you had to spend the rest of your life with the symptoms you have now, how would you feel about it?" The response categories are very dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, and very satisfied.

Progress rating—The participant is asked, "How would you rate your progress with your spine-related problem since you first enrolled in SPORT?" Response categories are major improvement, minor improvement, no change, minor worsening, and major worsening.

Self-perceived Health—The first question of the SF-36 which asks, "In general would you say your health is excellent, very good, good, fair, or poor."

Participants first completed the ODI, followed by the SF-36; the EQ-5D including VAS rating; a symptom satisfaction rating; progress rating; and the HUI.

Statistical Analyses

We summarized participant characteristics according to the four criteria: ODI, symptom satisfaction rating, progress rating, and self-perceived health rating. Mean change scores

(change scores) were calculated for each system from baseline to one year. We described the distribution of change scores using means, standard deviations and ranges. We summarized the distribution of health state classifications by dimension and level using percents.

We tested differences between change scores using signed rank tests. We assessed longitudinal validity by calculating Spearman correlation coefficients for change scores for system pairs and using tests for trend across changes in levels of each criterion measure. We evaluated floor and ceiling effects for each system by calculating the proportion of participants who received the highest and lowest possible scores at baseline and at one year. This analysis was repeated for the key dimensions of pain and physical function.

We calculated responsiveness statistics and 95% confidence intervals at one year for each system using distribution- and anchor-based methods.

We calculated distribution-based effect size and standardized response mean estimates as follows:

$$\text{Effect Size}[46] = \frac{\text{Mean}(\text{follow-up} - \text{baseline})}{SD_{\text{baseline}}}$$

$$\text{Standardized Response Mean (SRM)}[46] = \frac{\text{Mean}(\text{follow-up} - \text{baseline})}{SD_{\text{change score}}}$$

We calculated anchor-based Minimal Important Difference (MID) estimates and 95% confidence intervals according to four anchors: ODI, symptom satisfaction, progress rating, self-perceived health (SPH) rating. MID was calculated as the mean change for those who reported minimal important change according to each anchor. The scores of those who worsened were multiplied by -1 . [47, 48] Minimal important change was defined as one level of change from baseline to one year for symptom satisfaction and self-perceived health. For progress rating, minimal important change was defined as report of minimal improvement or minimal worsening at one year. SPORT sample size calculation was based on a 10-point change in ODI, and consistent with this and other work on important change for ODI, we used a 10–19 point change as the definition of minimal important change in this study. [43, 49]

All analyses were undertaken using STATA, version 9 (STATA Corporation, College Station, Texas).

Results

Data at one year were available for 1,000 participants whose mean age was 42 years (± 11) (Table 1). This was a highly educated population with the majority classifying their race as white. The majority of participants reported improved health based on all criterion measures.

A summary of mean change in health state values (change scores) is shown in Table 2. The largest mean change score was 0.40 for EQ-5D-UK, which was 3 times the mean change of 0.13 for eQWB. Standard deviations of the change scores were largest for EQ-5D-UK, followed by HUI3, EQ-5D-US, HUI2, SF-6D, and eQWB. Standard deviations were largest for the change scores, followed by baseline scores, and smallest for the 1-year scores.

Correlation between change scores as measured by Spearman coefficients ranged from 0.55 to 0.99 (Table 3). Not surprisingly, strongest correlations were noted between change scores from related systems, such as EQ-5D-UK and EQ-5D-US; HUI3 and HUI2; and SF6D and eQWB. Moderate to strong correlations were noted between change scores of all other systems.

When compared using sign rank tests, all change scores were significantly different from each other except EQ-5D-US and HUI2. All systems demonstrated linear trends and high correlations between change scores and change in levels of ODI, symptom satisfaction, progress rating, and self-perceived health (Figures 1a–d).

At one year, less than 1% of participants received the lowest possible score for each system, and 28% received the highest possible score for EQ-5D-UK and EQ-5D-US (Table 2). In contrast, HUI3 and HUI2 classified less than 10% at the ceiling, SF-6D defined 5%, and eQWB classified less than 1%. At baseline, each system classified a significant proportion of patients at the worst level for usual/physical function and pain. For the pain dimension, % at the floor was: 40% for EQ-5D; 29% for HUI3; 19% for HUI2; and 28% for SF-6D. For mobility/physical function, it was 3% for EQ-5D; 2% HUI3 and HUI2; and 14% for SF-6D. EQ-5D classified 25% at the floor for usual activities. At one year, there were large proportions at the best level (ceiling) for all dimensions. Specifically, in the mobility/physical function dimension, % at the ceiling was: 73% for EQ-5D; 82% for HUI3 and HUI2 ; and 22% for SF-6D. The proportions at the ceiling for pain were: 33 % for EQ-5D, 20% for HUI3 and HUI2; and 12% for SF-6D. All systems classified a smaller proportion at the floor at one year. Floor effects were also noted for EQ-5D in usual activities and pain/discomfort, and SF-6D in role limitations and vitality.

Table 4 summarizes responsiveness statistics. The estimated QWB score was most responsive, followed by the SF-6D. EQ-5D-UK was consistently the least responsive, although EQ-5D-US, HUI3 and HUI2 demonstrated similar or slightly less responsiveness. For example, the effect sizes for EQ-5D-UK and eQWB were 1.2 and 2.3 respectively. The Standardized Response Means (SRM) were 1.1 and 1.4 respectively.

Overall, MIDs were smaller for eQWB and SF-6D than for the other four systems. Values for EQ-5D-UK were approximately three to five times larger than those for eQWB. For example using ODI as the anchor, the MID estimate for EQ-5D-UK was 0.12, while the estimate for eQWB was 0.05.

Discussion

Our study is the first large longitudinal comparison of preference-weighted system performance in persons with confirmed diagnosis of IDH. Correlations between systems and tests of trend with external criteria support the notion that all systems were valid measures of HRQL. Estimates of effect size and standardized response mean indicates that all systems demonstrated the ability to measure change in key dimensions of HRQOL in this population of persons with spine disorders.

Considering the results of our validity tests together with the differences in mean change in health state values across systems, it is clear that there is no one system whose overall performance was superior to others. For example the superior responsiveness of eQWB and SF-6D, evidenced by the effect size and SRM estimates found in this study, confers an important advantage by enabling the detection of change with fewer study participants. Similarly, MID estimates indicated that EQ-5D-UK would require a larger magnitude of change to be considered clinically important compared to eQWB. However, the limited variation in scores upon which estimates of responsiveness are based has implications for policy applications. eQWB and to a lesser extent, SF-6D did not provide scores across the full range of health state values relative to the anchors of dead and perfect health. Our study findings were consistent with other comparisons that support overall validity of all systems, and somewhat better responsiveness for SF-6D, paired with potential overvaluation of lower

health states.[32, 50] Other studies indicate variations in performance across diagnoses and severity of health states.

Unique characteristics of systems

In the absence of a clearly superior system, the combination of unique strengths and limitations incorporated into preference-weighted health state classification systems presents difficult tradeoffs for researchers considering system choice. SF-6D demonstrated superior responsiveness and fewer ceiling problems for the pain and mobility dimensions. It is based on the longest of the surveys, with 36 questions, and covers several dimensions particularly relevant to persons with spine problems. Although the SF-36-derived approach may convey advantages in terms of responsiveness and ceiling effects, there is some indication that SF-6D may provide higher values for more severe health states and therefore may undervalue interventions.

The ease of administration is a key advantage of EQ-5D-UK and EQ-5D-US. However, floor and ceiling effects in dimensions highly relevant to spine disorders should be weighed against resource efficiency. EQ-5D-UK does not provide health state values between 0.88 and 1, and has been shown to provide lower mean health state values for health states compared to other systems.[14, 36]

The questionnaire for HUI2 and HUI3 is of intermediate length compared to SF-36 and EQ-5D, with fifteen questions. HUI3 incorporates dimensions not likely to be critical in the spine population, such as speech, vision, and hearing, and the dimensions covering mobility are limited to ambulation and dexterity. HUI3 has potential limitations in characterizing diminished mobility other than ambulation.[8] HUI2 includes mobility, and self-care dimensions important aspects of HRQOL for persons with spine disorders. Both HUI3 and HUI2 demonstrated limitations in characterizing change in mobility in our study.

Practical issues should also be considered in choosing a measurement system. The tradeoffs between resources required for survey administration, acceptability to participants, and measurement properties must be considered carefully. The systems considered in this study range from 5 to 36 questions, with various response levels. Depending on the research context, these may represent important differences.

Comparisons with Other Studies

Although all systems demonstrated very similar patterns of psychometric performance across all criteria, some important differences emerged that can be compared to prior research. Although responsiveness statistics indicated acceptable performance for all systems, eQWB, and SF-6D demonstrated the highest responsiveness, as indicated by larger effect size and SRM and smaller MIDs. The EQ-5D-UK had the lowest responsiveness of the measures. Other studies of the EQ-5D-UK and SF-6D in persons with various conditions have found similar. [34, 51–54] Walters and Brazier found slightly better responsiveness for SF-6D than EQ-5D in the results of combined analysis of data from 11 cohorts.[48] In the same study, the SRM and effect size were larger for SF-6D than for EQ-5D among patients with unspecified back pain. Longworth and Bryan found that SF-6D was limited in capturing change for severe health states but was more responsive among better health states compared to EQ-5D-UK among liver transplant patients.[30] In contrast, Conner-Spady found slightly better responsiveness for EQ-5D-UK compared to HUI3 and SF-6D among rheumatology patients stratified by change status.[24]

We found that responsiveness statistics for HUI3, and HUI2, were similar to or slightly better than those for EQ-5D-UK and US and slightly worse than those for SF-6D and eQWB. Studies conducted among persons with stroke, epilepsy, and heart disease have

reported similar patterns [20, 21, 55] In contrast, Feeny et al. found better responsiveness for HUI3 and HUI2 than SF-6D among patients undergoing hip replacement.[22] HUI3 demonstrated slight advantages over HUI2 in responsiveness among patients undergoing breast reduction surgery.[56]

Responsiveness statistics, including MIDs are important for study planning and for interpretation of changes in each system among patients with symptoms related to IDH. Estimates of effect size, SRM, and MID were generally larger in magnitude in our study than in other studies. [20–22, 24, 30, 34, 51–55] This can be explained by the large functional health status changes in our population over the study period. Our MID results for those who reported minimal change were 0.08 for SF-6D and 0.15 for EQ-5D-UK, meaning that a mean change of 0.08 in a clinical study using SF-6D would correspond with the lowest threshold for important change from the patient perspective using progress rating as the criterion. Alternatively, when judging the magnitude of change reported in clinical studies, a mean difference between treatment arms of 0.08 would indicate clinically meaningful difference using SF-6D. However, using EQ-5D-UK, the threshold would be 0.15.

Similarly, deficits in coverage for systems indicated by floor or ceiling problems have very important implications for system performance in measuring change over time. No ceiling effect was noted at baseline in our study. Consistent with previous studies conducted using data from persons with health conditions and from general population samples, large ceiling effects were noted in our study for the overall preference score for EQ-5D-UK and EQ-5D-US at one-year follow-up. [14, 18, 36, 53, 57] Smaller, but potentially significant ceiling effects were noted for HUI3, HUI2, and SF-6D. These results would indicate that EQ-5D-UK and US may have difficulty characterizing change for long-term outcomes compared to other instruments.

Perhaps of greatest significance was the remarkable proportion of patients at the ceiling on all systems for the key dimensions targeted by treatment. All systems classified significant proportions of participants at the floor or the ceiling of the pain and mobility dimensions at baseline or follow-up. These dimensions are particularly relevant for measuring effects of treatment this population. Floor effects were evident at baseline in the pain dimension for EQ-5D-UK, EQ-5D-US, HUI3, HUI2, and SF-6D Our review of the literature found that SF-6D demonstrated floor effects, and a limited range of available scores. This is consistent with other studies of SF-36 in this population.[15, 27, 30, 58]

Ceiling effects were evident in the mobility dimension at baseline for EQ-5D-UK, EQ-5D-US, HUI3, and HUI2, and in mobility and pain dimensions for all systems at one year. Ceiling effects were greater for EQ-5D-UK and EQ-5D-US than other systems in the pain dimension. In contrast, ceiling effects were greater for HUI3 and HUI2 in the mobility dimension. Feeny et al. suggested that HUI3 may be limited in detecting changes in mobility that do not involve the hands. [8] Our findings are consistent with this concern. However, HUI2 performed similarly to HUI3 in spite of describing mobility in broader terms.

Although psychometric evaluations are fundamental in establishing the measurement characteristics of preference-weighted systems, it is critical to assess validity in the context of their application for policy decision making. To address this question, we compared estimates of mean score change, or mean change in health state value, since this estimate is fundamental to QALY calculation. Except for EQ-5D-US and HUI2, we found that systems produced significantly different estimates of mean change in health state value. EQ-5D-UK produced the largest estimate of mean change, followed by HUI3, HUI2 and EQ-5D-US, SF-6D, and finally eQWB. Other studies have found differences in head-to-head

longitudinal comparisons.[20–34, 51, 54–56, 59] Similarly, these studies reported that EQ-5D-UK estimates were generally largest, followed by HUI3, HUI2, and finally SF-6D. These patterns are generally consistent with the results of cross-sectional comparisons.

Although comparisons of mean health state values were more common, we identified comparisons of QALYs or ICERs obtained using relevant systems in the published literature. Pickard et al. found that QALY differences calculated using EQ-5D-UK or HUI3 were two times larger than those obtained using SF-6D or HUI2.[20] Tosteson et al. reported ICERs for EQ-5D-UK and SF-6D in their cost-effectiveness analysis of surgery relative to non-operative treatment for persons with spinal stenosis with and without degenerative spondylolisthesis.[60] The ICER (95% CI) for spinal stenosis using the EQ-5D-UK was \$77,600 (\$49,564–\$120,042) compared to \$93,400 (\$59,205–\$143,660) using the SF-6D. The ICER for spinal stenosis with degenerative spondylolisthesis was \$115,600 (\$90,839–\$144,863) using the EQ-5D-UK compared to \$172,500 (\$132,178–\$221,930) using SF-6D. Van den Hout reported change in QALYs for EQ-5D-UK, EQ-5D-US, and SF-6D in their cost-utility analysis of early surgery versus prolonged conservative care among patients with sciatica from IDH.[61] Their results were consistent with our findings: that EQ-5D-UK provided the largest change in health state values, and smallest cost-utility ratio, followed by EQ-5D-US, and SF-6D. The gains in QALYs were: 0.044 (95%CI: 0.005 to 0.083); 0.032 (0.005 to 0.059), the SF-6D of 0.024 (0.003 to 0.046). Another study investigating acupuncture compared to usual care for nonspecific back pain reported very similar ICERS using EQ-5D-UK and SF-6D.[62] Joore et al found differences in ICERs and the probabilities of acceptability for the ICERs across five conditions using EQ-5D-UK and SF-6D.[63] Specifically, higher probabilities of acceptability were found using EQ-5D-UK for milder health conditions and using SF-6D for more severe health conditions. This study highlights the need to assess the performance of preference-weighted health state classification systems for specific conditions.

The results of our study indicate that using a regression model to “map” from SF-36-based health states to the QWB score for persons with intervertebral disc herniation may be a reasonable approach. The eQWB performed well in psychometric validity tests, but provided the lowest estimation of mean change in health state value, and very little variation in preference estimates. Kaplan et al. reported similar results among patients with arthritis[29] Although the psychometric properties of SF-6D and eQWB may be fairly similar, we recognize the additional steps of incorporating direct valuations as an advantage of the SF-6D over the “mapped” estimates produced by eQWB. Furthermore, our results indicate that eQWB may be more likely to produce qualitatively different cost-utility results than alternative systems. However, interest in developing methods to estimate health state values from existing HRQOL data appears to be increasing, and their relative performance should be investigated. It should be noted that the performance of the eQWB is dependent on the characteristics of SF-36 health state classification system, the Quality of Well Being Scale, and the regression model used to link the two. Our results indicate that the eQWB estimates may be used with some caution in this population, but this should be weighed carefully against the advantages of the SF-6D.

Limitations

As with any validation studies of HRQOL instruments, there is no gold standard for the performance of preference-weighted health state classification systems. Although it is possible to generate and test hypotheses about the behavior of the systems under known circumstances, some variation would be expected between the systems under examination and the measure used to test validity.

We used a longitudinal cohort design to calculate responsiveness statistics, including effect size. This is not the same computation as is used to characterize the effect of treatment compared to control. It may be argued that for the purpose of interpretation in clinical trials, estimation of the effect size of treatment is the most relevant calculation. Because all patients in this trial undergo some treatment, either surgical or non-surgical, and most were expected to improve, we measured observed change in the entire group. However, effect size is commonly used in applications similar to ours to address longitudinal validity.

Since the order of administration of instruments was not counterbalanced, we cannot rule out an order effect. However, because the questions in each instrument are of similar nature, it is doubtful that there would be significant learning or framing effects exhibited in this application.

In summary, this study provided information about the performance and interpretation of several of the most widely used preference-weighted health state classification systems. The evidence supports the notion that all systems are measuring the same construct, but each has unique characteristics that should be considered when choosing a system. We found evidence that all systems demonstrate validity in this population, with some caveats. All systems demonstrated evidence of ceiling or floor effects for key dimensions relevant to spine disorders. In the context of cost-effectiveness analysis, we found that change scores were significantly different except for EQ-5D-US and HUI2. Change scores were largest for EQ-5D-UK, followed by HUI3, HUI2, EQ-5D-US, SF-6D, and finally eQWB. Such differences indicate that care should be taken when interpreting cost-utility analyses from different systems. Researchers choosing a system should carefully consider the characteristics of each system relative to study goals.

Acknowledgments

Acknowledgement of Support: The authors would like to acknowledge funding from the following sources: Grant Number F32HD056763 from the National Institute of Child Health And Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Child Health And Human Development or the National Institutes of Health. Support for this research was provided by the National Institute of Arthritis and Musculoskeletal and Skin Diseases (U01-AR45444-01A1 and P60-AR048094-01A1) and the Office of Research on Women's Health, the National Institutes of Health, and the National Institute of Occupational Safety and Health, the Centers for Disease Control and Prevention and a New Investigator Fellowship Training Initiative grant from the Foundation for Physical Therapy.

References

1. Bloom BS. Use of formal benefit/cost evaluations in health system decision making. *American Journal of Managed Care*. 2004 May; 10(5):329–335. [see comment]. [PubMed: 15152703]
2. Dickson M, Hurst J, Jacobzone S. Survey of pharmacoeconomic assessment activity in eleven countries: OECD. 2003
3. Gold, MR.; Siegel, JE.; Russell, LB.; Weinstein, MC. *Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press; 1996.
4. Kaplan RM, Anderson JP. A general health policy model: update and applications. *Health Services Research*. 1988 Jun; 23(2):203–235. [PubMed: 3384669]
5. Brooks R. EuroQol: the current state of play. *Health Policy*. 1996; 37(1):53–72. [PubMed: 10158943]
6. Dolan P. Modelling valuations for EuroQol Health States. *Medical Care*. 1997 Nov; 35(11):1095–1108. [PubMed: 9366889]
7. Boyle MH, Furlong W, Feeny D, Torrance GW, Hatcher J. Reliability of the Health Utilities Index--Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. *Quality of Life Research*. 1995; 4(3):249–257. [PubMed: 7613535]

8. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. Pharmacoeconomics. 1995; 7(6):490–502. [PubMed: 10155335]
9. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions. Health Utilities Index. Pharmacoeconomics. 1995; 7(6):503–520. [PubMed: 10155336]
10. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics. 2002 Mar; 21(2):271–292. [PubMed: 11939242]
11. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. Journal of Clinical Epidemiology. 1998 Nov; 51(11):1115–1128. [PubMed: 9817129]
12. Richardson G, Manca A. Calculation of quality adjusted life years in the published literature: a review of methodology and transparency. Health Economics. 2004; 13:1203–1210. [PubMed: 15386669]
13. Barton GR, Bankart J, Davis AC. A comparison of the quality of life of hearing-impaired people as estimated by three different utility measures. International Journal of Audiology. 2005; 44:157–163. [PubMed: 15916116]
14. Belanger, A.; Berthelot, J-M.; Guimond, E.; Houle, C. A Head-to-Head Comparison of Two Generic Health Status Measures in the Household Population: McMaster Health Utilities Index (Mark 3) and the EQ-5D. Ottawa: Statistics Canada, Health Analysis and Modelling Group; 2000. Final Revision April 2000.
15. Brazier, J.; Roberts, J.; Tsuchiya, A. A Comparison of the EQ-5D and SF-6D Across Seven Patient Groups; Proceedings of the 18th Plenary Meeting of the Euroqol Group; 2001. p. 9-31.2001
16. Espallargues M, Czoski-Murray CJ, Bansback NJ, Carlton J, Lewis GM, Hughes LA, et al. The impact of age-related macular degeneration on health status utility values. Investigative Ophthalmology & Visual Science. 2005 Nov; 46(11):4016–4023. [PubMed: 16249475]
17. Luo N, Johnson JA, Shaw JW, Feeny D, Coons SJ. Self-reported health status of the general adult U.S population as assessed by the EQ-5D and Health Utilities Index. Medical Care. 2005; 43(11): 1078–1086. [PubMed: 16224300]
18. Marra CA, Esdaile JM, Guh D, Kopec JA, Brazier JE, Koehler BE, et al. A comparison of four indirect methods of assessing utility values in rheumatoid arthritis. Medical Care. 2004 Nov; 42(11):1125–1131. [PubMed: 15586840]
19. McDonough CM, Grove MR, Tosteson TD, Lurie JD, Hilibrand AS, Tosteson AN. Comparison of EQ-5D, HUI, SF-36-derived societal health state values among spine patient outcomes research trial (SPORT) participants. Quality of Life Research. 2005 Jun; 14(5):1321–1332. [PubMed: 16047507]
20. Pickard AS, Johnson JA, Feeny DH. Responsiveness of generic health-related quality of life measures in stroke. Quality of Life Research. 2005 Feb; 14(1):207–219. [PubMed: 15789955]
21. Hatoum HT, Brazier JE, Akhras KS. Comparison of the HUI3 with the SF-36 preference based SF-6D in a clinical trial setting. Value in Health. 2004 Sep-Oct; 7(5):602–609. [PubMed: 15367255]
22. Feeny D, Wu L, Eng K. Comparing Short Form 6D, Standard Gamble, and Health Utilities Index Mark 2 and Mark 3 utility scores: Results from total hip arthroplasty patients. Quality of Life Research. 2004; 13:1659–1670. [PubMed: 15651537]
23. Conner-Spady, B.; Suarez-Almazor, ME. A Comparison of preference-based health status tools in patients with musculoskeletal disease; 18th Plenary Meeting of the EuroQol Group; 2001. p. 235-245.2001
24. Conner-Spady B, Suarez-Almazor ME. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. Medical Care. 2003 Jul; 41(7):791–801. [PubMed: 12835603]
25. Conner-Spady, B.; Voaklander, DC.; Suarez-Almazor, ME. The effect of different EuroQol weights on potential QALYs gained in patients with hip and knee replacement; 17th Plenary Meeting of the EuroQol Group; 2000. p. 127-137.2000
26. Bosch JL, Hunink M. Comparison of the Health Utilities Index Mark 3 (HUI3) and the EuroQol EQ-5D in patients treated for intermittent claudication. Quality of Life Research. 2000; 9:591–601. [PubMed: 11236850]

27. Suarez-Almazor M, C K, Johnson J, Skeith K, D V. Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic and preference-based instruments. *Rheumatology*. 2000; 39:783–790. [PubMed: 10908699]
28. Bosch JL, Halpern EF, Gazelle GS. Comparison of preference-based utilities of the Short-Form 36 Health Survey and Health Utilities Index before and after treatment of patients with intermittent claudication. *Medical Decision Making*. 2002 Sep-Oct;22(5):403–409. [PubMed: 12365482]
29. Kaplan R, Groessl EJ, Sengupta N, Sieber WJ, Ganiats TG. Comparison of measured utility scores and imputed scores from the SF-36 in patients with rheumatoid arthritis. *Medical Care*. 2005 Jan; 43(1):79–87. [PubMed: 15626937]
30. Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Economics*. 2003 Dec; 12(12):1061–1067. [PubMed: 14673814]
31. Neumann PJ, Sandberg EA, Araki SS, Kuntz KM, Feeny D, Weinstein MC. A comparison of HUI2 and HUI3 utility scores in Alzheimer's disease. *Medical Decision Making*. 2000 Oct-Dec; 20(4):413–422. [PubMed: 11059474]
32. Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. *Annals of Medicine*. 2001 Jul; 33(5):358–370. [PubMed: 11491195]
33. Holland R, Smith RD, Harvey I, Swift L, Lenaghan E. Assessing quality of life in the elderly: a direct comparison of the EQ-5D and AQoL. *Health Economics*. 2004 Aug; 13(8):793–805. [PubMed: 15322991]
34. Stavem K, Froland SS, Hellum KB. Comparison of preference-based utilities of the 15D, EQ-5D and SF-6D in patients with HIV/AIDS. *Quality of Life Research*. 2005 May; 14(4):971–980. [PubMed: 16041894]
35. Hollingworth W, Deyo RA, Sullivan SD, Emerson SS, Gray DT, Jarvik JG. The practicality and validity of directly elicited and SF-36 derived health state preferences in patients with low back pain. *Health Economics*. 2002; 11(1):71–85. [PubMed: 11788983]
36. Macran S, Weatherly H, Kind P. Measuring population health: a comparison of three generic health status measures. *Medical Care*. 2003; 41(2):218–231. [PubMed: 12555050]
37. Birkmeyer NJ, Weinstein JN, Tosteson AN, Tosteson TD, Skinner JS, Lurie JD, et al. Design of the Spine Patient outcomes Research Trial (SPORT). *Spine*. 2002; 27(12):1361–1372. [PubMed: 12065987]
38. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Medical Care*. 2005 Mar; 43(3):203–220. [PubMed: 15725977]
39. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and Single-Attribute Utility Functions for the Health Utilities Index Mark 3 System. *Medical Care*. 2002; 40(2):113–128. [PubMed: 11802084]
40. Tosteson AN. Preference-based health outcome measures in low back pain. *Spine*. 2000 Dec 15; 25(24):3161–3166. [PubMed: 11124732]
41. Tosteson AN, Hammond CS. Quality-of-life assessment in osteoporosis: health-status and preference-based measures. *Pharmacoeconomics*. 2002; 20(5):289–303. [PubMed: 11994039]
42. Fryback DG, Lawrence WF, Martin PA, Klein R, Klein BE. Predicting Quality of Well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study. *Medical Decision Making*. 1997; 17(1):1–9. [PubMed: 8994146]
43. Fairbank JC, Pynsent PB. The Oswestry Disability Index. *Spine*. 2000 Nov 15; 25(22):2940–2952. discussion 52. [PubMed: 11074683]
44. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy*. 1980; 66(8):271–273. [PubMed: 6450426]
45. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. *Journal of Clinical Epidemiology*. 2001 Dec; 54(12):1204–1217. [see comment]. [PubMed: 11750189]
46. Terwee C, Dekker F, Wiersinga W, Prummel M, Bossuyt P. On assessing responsiveness of health related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research*. 2003; 12:349–362. [PubMed: 12797708]

47. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*. 1989 Dec; 10(4):407–415. [PubMed: 2691207]
48. Walters SJ, Brazier J. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Quality of Life Research*. 2005; 14:1523–1532. [PubMed: 16110932]
49. Bombardier C, Hayden J, Beaton DE. Minimal clinically important difference. Low back pain: outcome measures. *Journal of Rheumatology*. 2001 Feb; 28(2):431–438. [PubMed: 11246692]
50. O'Brien BJ, Spathe M, Blackhouse G, Severens JL, Dorian P, Brazier J. A view from the bridge: agreement between the SF-6D utility algorithm and the Health Utilities Index. *Health Economics*. 2003; 12:975–981. [PubMed: 14601159]
51. Lamers LM, Meerdling WJ, Severens JL, Brouwer WB. The relationship between productivity and health-related quality of life: an empirical exploration in persons with low back pain. *Quality of Life Research*. 2005 Apr; 14(3):805–813. [PubMed: 16022073]
52. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. *Social Science & Medicine*. 2005 Apr; 60(7):1571–1582. [PubMed: 15652688]
53. Oga T, Nishimura K, Tsukino M, Sato S, Hajiro T, Mishima M. A comparison of the responsiveness of different generic health status measures in patients with asthma. *Quality of Life Research*. 2003; 12:555–563. [PubMed: 13677500]
54. van Stel HF, E B. Comparison of the SF-6D and the EQ-5D in patients with coronary heart disease. *Health and Quality of Life Outcomes*. 2006; 4(20)
55. Langfitt J, Vickrey B, McDermott M, Messing S, Berg A, Spencer S, et al. Validity and responsiveness of generic preference-based HRQOL instruments in chronic epilepsy. *Quality of Life Research*. 2006; 15:899–914. [PubMed: 16721649]
56. Thoma A, Sprague S, Veltri K, Duku E, Furlong W. Methodology and measurement properties of health-related quality of life instruments: A prospective study of patients undergoing breast reduction surgery. *Health and Quality of Life Outcomes*. 2005; 3(44)
57. Bharmal M, Thomas J III. Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. *Value in Health*. 2006; 9(4):262–271. [PubMed: 16903996]
58. Taylor SJ, Taylor AE, Foy MA, Fogg AJ. Responsiveness of common outcome measures for patients with low back pain. *Spine*. 1999; 24(17):1805–1812. [PubMed: 10488511]
59. Sherbourne C, Unutzer J, Schoenbaum M, Duan N, Lenert L, Sturm R, et al. Can utility-weighted health-related quality-of-life estimates capture health effects of quality improvement for depression? *Medical Care*. 2001; 39(11):1246–1259. [PubMed: 11606878]
60. Tosteson AN, Lurie JD, Tosteson TD, Skinner JS, Herkowitz H, Albert T, et al. Surgical treatment of spinal stenosis with and without degenerative spondylolisthesis: cost-effectiveness after 2 years. *Annals of Internal Medicine*. 2008 Dec 16; 149(12):845–853. [see comment]. [PubMed: 19075203]
61. van den Hout WB, Peul WC, Koes BW, Brand R, Kievit J, Thomeer RT. Prolonged conservative care versus early surgery in patients with sciatica from lumbar disc herniation: cost utility analysis alongside a randomised controlled trial. *British Medical Journal*. 2008
62. Thomas, KJ.; MacPherson, H.; Ratcliffe, J.; Thorpe, L.; Brazier, J.; Campbell, M., et al. *Health Technology Assessment*. Vol. 9. England: Winchester; 2005 Aug. Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain.
62. Joore M, Brunenberg D, Nelemans P, Wouters E, Kuijpers P, Honig A, et al. The Impact of Differences in EQ-5D and SF-6D Utility Scores on the Acceptability of Cost–Utility Ratios: Results across Five Trial-Based Cost–Utility Studies. *Value in Health*. 2010 Mar-Apr; 13(2):222–229. [PubMed: 19878492]

1a Change in the Oswestry Disability Index scores are presented in 10 point intervals. For consistency of interpretation, ODI was coded so that negative values represent worsening. Due to small numbers of participants, intervals at the extremes of the range were collapsed.

1b Symptom satisfaction response levels include: very satisfied; somewhat satisfied; neutral; somewhat dissatisfied; and very satisfied. Due to the limited number of participants reporting worsening, levels of worsened symptom satisfaction were combined.

1c Progress Rating. Progress compared to baseline was directly rated by participants among these levels: major improvement; minor improvement; no change; minor worsening; and major worsening.

1d Self-Perceived Health Rating. Participants rated current health from among the following categories: excellent, very good, good, fair, poor. Due to limited number of participants reporting more than one level of change, categories were combined for improvement or worsening of two levels or greater.

Figure 1.

a–d Trends in mean change in health state value with levels of progress rating and change in ODI score, symptom satisfaction, and self-perceived health rating.

Table 1

Summary of characteristics of longitudinal study participants for SPORT participants with IDH.

	n=1,000
Age	
Mean±SD	42.2±11.3
<50 years	75.8
50–64 years	19.7
65 years and over	4.5
Female Sex	43.2
Body Mass Index 29	35.6
Enrolled in RCT	38.8
Ethnicity: Hispanic	2.8
Race	
Black/African American	5.5
White	87.9
Other (including unknown)	6.6
Education: College graduate	48.6
Married/living with a significant other	71.7

Table 2

Summary statistics at 1 year for preference-weighted systems among 1,000 SPORT participants with IDH

	Health State Value at 1 year (SD)	Mean Change Health State Value (SD)	% Ceiling
System			
EQ-5D-UK	0.76 (0.24)	0.40 (0.38)	28.2
EQ-5D-US	0.82 (0.17)	0.27 (0.25)	28.2
HUI3	0.78 (0.23)	0.34 (0.29)	6.2
HUI2	0.84 (0.16)	0.28 (0.23)	7.9
SF-6D	0.76 (0.15)	0.21 (0.16)	4.8
eQWB	0.71 (0.08)	0.13 (0.09)	0.8

One person scored at the floor for overall score on all systems (0.10% at floor).

\$watermark-text

\$watermark-text

\$watermark-text

Table 3
Spearman rank correlations between change scores for preference-weighted systems

	EQ-5D-US	HUI3	HUI2	SF-6D	eQWB
EQ-5D-UK	0.99	0.65	0.67	0.62	0.57
EQ-5D-US		0.65	0.67	0.62	0.57
HUI3			0.77	0.69	0.57
HUI2				0.63	0.55
SF-6D					0.80

Table 4

Responsiveness statistics for 1,000 SPORT participants with IDH

	Distribution-Based Statistics (95%CI)		Minimal Important Difference Estimates (95%CI)			
	Effect Size	SRM ^a	ODI	Satisfaction	Progress	Self-Perceived Health
System			n=172	n=170	n=194	n=403
EQ5D-UK	1.20 (1.13, 1.27)	1.06 (1.00, 1.13)	0.12 (0.09, 0.16)	0.25 (0.20, 0.30)	0.15 (0.10, 0.20)	0.14 (0.09, 0.19)
EQ-5D-US	1.23 (1.16, 1.30)	1.07 (1.01, 1.13)	0.08 (0.06, 0.11)	0.17 (0.13, 0.20)	0.10 (0.07, 0.13)	0.10 (0.06, 0.13)
HUI3	1.25 (1.18, 1.32)	1.17 (1.10, 1.23)	0.11 (0.08, 0.15)	0.21 (0.16, 0.25)	0.15 (0.11, 0.19)	0.13 (0.09, 0.18)
HUI2	1.24 (1.18, 1.31)	1.18 (1.11, 1.24)	0.09 (0.065, 0.12)	0.18 (0.15, 0.21)	0.11 (0.08, 0.14)	0.09 (0.05, 0.12)
SF-6D	1.89 (1.80, 1.98)	1.34 (1.28, 1.40)	0.09 (0.07, 0.11)	0.11 (0.09, 0.13)	0.08 (0.06, 0.10)	0.08 (0.06, 0.11)
eQWB	2.32 (2.22, 2.42)	1.42 (1.36, 1.49)	0.05 (0.04, 0.06)	0.06 (0.05, 0.07)	0.03 (0.02, 0.04)	0.05 (0.04, 0.07)

^aSRM = Standardized Response Mean